IJCSC Volume 5 • Number 1 March-Sep 2014 pp.57-60 ISSN-0973-7391

Semantic and Sentiment Analysis

Prof Shivani Desai, Priyank Bhatt(11BIT044), Vraj Solanki(11BIT059)
Nirma University, Ahmedabad, India
Shivani.desai@nirmauni.ac.in,11bit044@nirmauni.ac.in, 11bit059@nirmauni.ac.in

***Abstract*—**Sentiment and Semantic analysis is a very powerful tool in today's internet. It is very important to find out the correct context and sense in which a particular sentence has been written on the internet because there is no physical contact to find out the meaning of the sentence. A number of methods and techniques are followed in order to classify the defined statement as positive or negative. This classification helps to actually find out the context of a sentence remotely. This paper aims at surveying a number of such algorithms, methods and techniques to classify any sentence as positive, negative or neutral and also discuss the issues related to each method faced during implementation and execution. The essential issues in sentiment analysis are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) opinions toward the subject and how efficiently and correctly sentences are classified.

Keywords- Sentimental analysis, SVM.

I. INTRODUCTION

Sentiment Analysis (SA) (Nasukawa and Yi, 2003; Yi et al., 2003)is a task to recognize writers' feelings as expressed in positive or negative comments, by analyzing unread ably large numbers of documents. Extensive syntactic patterns enable us to detect sentiment expressions and to convert them into semantic structures with high precision, as reported by Kanayama et al. (2004).

It is a very important method in today's world where the majority of our work is carried on internet be it communicating with clients reading news, blogs placing reviews about a company, product or person. So it becomes very important for us to detect the exact meaning of the sentence written or else it may lead to disastrous (in many cases completely opposite) understanding of the issue. The essential issues in sentiment analysis are to identify how sentiments are expressed in texts and whether the expressions indicate positive (favorable) or negative (unfavorable) perspective toward the subject.

In current scenario approaches to extract sentiments associated with polarities of positive or negative for specific subjects from a document, instead of classifying the whole document into positive or negative are used so a large amount of information is available from a single document.

Many of their applications aim to classify the whole document into positive or negative toward a subject of the document that is specified either explicitly or implicitly For example, the classification of a movie review into positive or negative, assumes that all sentiment expressions in the review represent sentiments directly toward that movie, and expressions that violate this assumption (such as a negative comment about an actor even though the movie as a whole is considered to be excellent) confuse the judgment of the classification. On the contrary, by analyzing the relationships between sentiment expressions and subjects, we can make in-depth analyses on what is favored and what is not. Thus such techniques to detect favorable and unfavorable opinions toward specific subjects within large numbers of documents offer enormous opportunities for various applications. It would provide powerful functionality for competitive analysis, marketing analysis, and detection of unfavorable rumors for risk management.

For example, enormous sums are being spent on customer satisfaction surveys and their analysis. Yet, the effectiveness of such surveys is usually very limited in spite of the amount of money and effort spent on them, both because of the sample size limitations and the difficulties of making effective questionnaires. Thus there is a natural desire to detect and analyze favorability within online documents such as Web pages, chat rooms, and news articles, instead of making special surveys with questionnaires. Humans can easily recognize natural opinions among such online documents. In addition, it might be crucial to monitor such online documents, since they sometimes influence public opinion, and negative rumors circulating in online documents may cause critical problems for some organizations.

Let us take an example to understand the actual application of sentiment analysis: "Product A is good but expensive." This statement contains a combination of two statements:

"Product A is good"

"Product A is expensive"

We think it's easy to agree that there is one statement, Product A is good, that indicates a favorable sentiment, and there is another statement, Product A is expensive, that indicates an unfavorable sentiment. Thus, instead of analyzing the favorability of the whole context, we try to extract each statement on favorability, and present them to the end users so that they can use the results according to their application requirements. Thus, sentiment analysis involves identification of:

- Sentiment expressions.
- Polarity and strength of the expressions.
- Their relationship to the subject.

These elements are interrelated. For example, in the sentence, "XXX beats YYY", the expression "beats" denotes a positive sentiment toward XXX and a negative sentiment toward YYY [1].

II. ANALYSIS OF ALGORITHMS

1) Finding key words from the sentence and classify the sentence as positive, negative or neutral POS method [1].
For POS (polarity of sentiments) tagging, Markov-model-based tagger is used essentially [5]. This tagger assigns a part of speech to text tokens based on the distribution probabilities of candidate POS labels for each word and the probability of a POS transition extracted from a training corpus. A manually annotated corpus of Wall Street Journal articles from the Penn Treebank Project was used as the training corpus by (Nasukawa and Yi, 2003). For these experiments, the tagger was configured to treat unknown words (i.e. those not seen in the training corpus, and excluding numbers) as nouns. The tagger uses a lexical look-up component, which offers sophisticated inflectional

analysis for all known words. Thus the tasks performed during the process are:

- The polarity of the sentiments
- The sentiment expressions that are applied, (identifying phrase boundaries i.e. sentence of impact and application).
- The phrases that contain the sentiment expressions are identified for given subject term.

As for a simple example let's take sentences:

"X provides a good working environment." --- [a]

"X provides a bad working environment." ---- [b]

Where "X" is a subject term with favorable and unfavorable sentiment, in statements [a] and [b] provided that "a good working environment" and "a bad working environment" are favorable and unfavorable, respectively.

Consider one more example: "X prevents trouble."

In which "X" is a subject term receiving favorable sentiment, and "trouble" is a sentiment term for unfavourability. So there the word is classified improperly because just looking the verb trouble improperly classifies the sentence.

Thus in this method terms are directly classified into positive negative and neutral.

ISSUES WITH THE METHOD:

1. Classification of terms just on the basis of the word is very primitive and trivial method.
2. In majority of the situations the words are classified wrongly because the meaning of the word depends only upon the context in which it is used. For example: "It's difficult to take a bad picture with this camera.[8] The algorithm extracts bad---picture (a bad picture) This is a positive statement for the camera, and it's not relevant to extract this "bad picture" as a negative sentiment.
3. Sometimes the sentences are just used as input condition to deciding statements. But they cannot be classified as well. For example: "Also the battery went dead while at Animal Kingdom and one feature I used to like about the Olympus is that if the recharge-able batteries went dead you could just pop some AA's in and still get your pictures." The algorithm performs: battery (the battery)---go (went)---dead (dead) Here the incident that "the battery went dead" is described as a normal event instead of product failure.
4. The same word can be used positively as well as negatively, and this difference could only be told by looking at the context. For example: prevent, risk ("put something at risk" may be favorable when the "something" is unfavorable such as the case of "hackers).
5. Entries in the table have to be made manually that is very tedious task.
6. The tagger may not be configured to handle unknown words in sentence. For example: ('to be free from vignette'), vignette not present in dictionary so can't be classified.

2) Tracking the reference frequencies of adjectives with positive and negative references to the sentences.[2] . Initially a small candidate seed list of positive and negative words is taken which is expanded into full sentiment lexicons using path-based analysis of synonym and antonym sets in WordNet. Sentiment-alternation hop counts are used to determine the polarity strength of the candidate terms and eliminate the ambiguous terms. The algorithm takes in consideration 2 values: polarity and subjectivity.

- Polarity: Is the sentiment associated with the entity positive or negative?
  - World_polarity: world polarity =(evaluate world polarity using sentiment data for all entities for the entire time period).
    positive sentiment references /total sentiment references.
  - Entity_polarity(entity polarity$_i$ using sentiment data for that day day$_i$ only)
    positive sentiment references$_i$/ total sentiment references$_i$.
- Subjectivity: The subjectivity time series reflects the amount of sentiment an entity is associated with, regardless of whether the sentiment is positive or negative..
  - world subjectivity: We evaluate world subjectivity using sentiment data for all entities for the entire time period.
    total sentiment references/ total references (For e.g: I cannot meet girl who is not beautiful).
  - entity subjectivity$_i$: We evaluate entity subjectivity$_i$ using sentiment data for that day (day$_i$) only.
    total sentiment references$_i$/ total references$_i$.(eg I want to meet a beautiful girl).

Example to differentiate subjectivity and polarity:

"No bad picture can be clicked with this camera."

If only bad is seen that is the polarity of word then it would denote a wrong meaning, but when the context is the subjectivity of word is seen it becomes clear that it is not a negative reference.

ISSUES WITH THE METHOD:

1. The method is very basic and trivial. It includes lot of calculation for each word taking in consideration the entire meaning of the sentence and classifying the word depending upon the subjectivity.
2. The generation of the token dictionary becomes tedious. It also considers mainly the average meaning of the word from all the sentences that have the occurrences of the word. But it may be possible that the classifier has always got those sentences as an input where the meaning of a word was always (specific i.e. either positive or negative). So it would averagely and mainly classify the word accordingly and when the same word with opposite meaning in a sentence comes to classifier it fails to properly detect the meaning.

3) Polarity assignment using polar atoms [3].

To assign a polarity to each proposition, polar atoms in the lexicon are compared to the proposition. A polar atom consists of polarity, verb or adjective, and optionally, its arguments. For example beautiful is a simple polar atom, where no argument is specified.

Simple: 'to be beautiful'

This atom matches any proposition which has beautiful in it.

Complex: 'to lack ← attraction-ACC'

I.e. to lack (argument specifies a specific quality for eg not a good orator).

A polarity is assigned if there exists a polar atom for which verb/adjective and the arguments coincide with the proposition, and otherwise no polarity is assigned. The opposite polarity of the polar atom is assigned to a proposition which has the negative feature.

Coherent density: cd (d,L) = (Coherent)/(Polar). This indicates the ratio of polar clauses that appear in the coherent (same) context, among all of the polar clauses detected by the system. High cd means that sentiment expressions are more frequently used in that domain. For each candidate polar atom a, the total appearances f(a), and the occurrences in positive contexts p(a) and negative contexts n(a) are counted, based on the context of the adjacent clauses.

To determine which words are to be added in sentiment dictionary: In order to set general criteria, here we assume that a true positive polar atom a should have higher p(a)/f(a) than its average i.e. coherent density, cd (d,L+a), and also have higher p(a)/p(a)+n(a) than its average i.e. coherent precision, cp(d,L+a).

Assuming the binomial distribution, a candidate polar atom is adopted as a positive polar atom if both conditions are fulfilled.

i. $q > cd(d,L)$,
Where $\sum_{p(a)}^{k=0} {}_{f(a)}C_k\, q^k (1-q)^{f(a)-k} = 0.9$

ii. $r > cp(d,L)$ or $n(a) = 0$,
where $p(a) \sum k=0\ {}_{p(a)+n(a)}C_k r^k (1-r)^{p(a)+n(a)-k} = 0.9$.

[7]

We can assume cd(d,L+a) as cd(d,L), and cp(d,L+a) as cp(d,L) when L is large.

ISSUES WITH THE METHOD:

1. In the evaluation process, some interesting results were observed. For example, a negative atom ('to be free from vignette') was acquired in the digital cam- era domain. Even the evaluator who was familiar with digital cameras did not know the term ('vignette'), but after looking up the dictionary she labeled it as negative. Our learning method could pick up such technical terms and labeled them appropriately. So some of the words that were not present in the atom dictionary could not be classified positive or negative.
2. An evaluator assigned positive to: ('to have camera') in the mobile phone domain, but the acquired polar atom had the negative polarity. This was actually an insight from the recent opinions that many users want phones without camera functions.

4) Semantic orientation with PMI [4].

Here, the term semantic orientation (SO) (Hatzi- vassiloglou and McKeown, 2002) refers to a real number measure of the positive or negative sentiment expressed by a word or phrase. This approach is simple and surprisingly effective. Moreover, is not restricted to words of a particular part of speech, nor even restricted to single words, but can be used with multiple word phrases. In general, two word phrases conforming to particular part-of-speech templates representing possible descriptive combinations which are used. Once the desired value phrases have been extracted from the text, each one is assigned an SO value. The SO of a phrase is determined based upon the phrase's point wise mutual information (PMI) with the words "excellent" and "poor". PMI is defined by Church and Hanks (1989) as follows:

$pmi(w1,w2)=\log_{2}((p(w1 \text{ and } w2))/(p(w1).p(w2)))$

Where p(w1 and w2) is probability that p1 and p2 co occur. The SO for a word is its difference from the word excellent and the word poor. The probabilities are estimated by querying the AltaVista Advanced Search engine1 for counts. The search engine's "NEAR" operator, representing occurrences of the two queried words within ten words of each other in a text, is used to define co-occurrence. The final SO equation is

$SO(phrase)=\log_{2}((hits[phrase \text{ near 'excellent'}]$
$hits['poor'])/(hits[phrase \text{ near 'poor'}]hits['excellent']))$

This yields values above 0 for phrases with greater PMI with the word "excellent" and below zero for greater PMI with "poor". A SO value of zero would indicate a completely neutral semantic orientation.

5) Osgood semantic differentiation with WordNet.[4]

Further feature types are derived using the method of Kamps and Marx (2002) of using WordNet relationships to derive three values pertinent to the emotive meaning of adjectives. These values are derived by measuring the relative minimal path length (MPL) in WordNet between the adjective in question and the pair of words appropriate for the given factor.

For example: In the case of the evaluative factor (EVA), the comparison is between the MPL between the adjective and "good" and the MPL between the adjective and "bad".

The three values correspond to the:

i. Potency (strong or weak)
ii. Activity (active or passive)
iii. Evaluative (good or bad) factors introduced in Charles Osgood's Theory of Semantic Differentiation (Osgood et al., 1957).

Only adjectives connected by synonymy to each of the opposites are considered, each of which is given a value for each of the three factors referred to as POT, ACT and EVA. For the purposes of this re- search, each of these factors' values are averaged over all the adjectives in a text, yielding three real- valued feature values for the text, which will be added to the SVM model.

6) Topic proximity and syntactic-relation features. [4]

This approach shares the intuition of Natsukawa and Yi (2003) that sentiment expressed with regard to a particular subject can best be identified with reference to the subject itself. Collecting emotive con- tent from a text overall can only give the most general indication of the sentiment of that text towards the specific subject. In some application domains, it is known in advance what the topic is toward which sentiment is to be evaluated. The present approach allows for the incorporation of features which exploit this knowledge, where available. This is done by creating several classes of features based upon the semantic orientation values of phrases given their position in relation to the topic of the text. Although in opinion-based texts there is generally a single primary subject about which the opinion is favorable or unfavorable, it would seem that secondary subjects may also be useful to identify. The

primary subject of a book review, for example, is a book. However, the review's overall attitude to the author may also be enlightening, although it is not necessarily identical to the attitude towards the book. Likewise in a product review, the attitude to- wards the company which manufactures the product may be supportive. It is an open question whether such secondary topic information would be beneficial or harmful to the modeling task. This approach allows such secondary information to be incorporated, where available. In each record review, references (including co-reference) to the record being reviewed were tagged as 'THIS WORK' and references to the artist under review were tagged as 'THIS ARTIST'.

Thus the decision classes are:

1.) **Turney Value**: The average value of all value phrases' SO values for the text. Classification by this feature alone is not the equivalent of Turney's approach, since the present approach involves retraining in a supervised model.
2.) **In sentence with THIS WORK**: The average value of all value phrases which occur in the same sentence as a reference to the work being reviewed.
3.) **Following THIS WORK**: The average value of all value phrases which follow a reference to the work being reviewed directly, or separated only by the preposition.
4.) **Preceding THIS WORK**: The average value of all value phrases which precede a reference to the work being reviewed directly, or separated only by the copula or a preposition.
5.) **In sentence with THIS ARTIST**: As above, but with reference to the artist.
6.) **Following THIS ARTIST:** As above, but with reference to the artist.
7.) **Preceding THIS ARTIST**: As above, but with reference to the artist.

ISSUES WITH THE METHOD:

1. Some improvement could be gained by adding domain context
2. Favorability content depends to some extent on their context: unpredictable is generally positive when describing a movie plot, and negative when describing an automobile or a politician. Likewise, such terms as "devastating" might be generally negative, but in the context of music or art may imply an emotional engagement which is usually seen as positive. Likewise, although "excellent" and "poor" as the poles in assessing this value seems somewhat arbitrary.
3. One problem with limiting the domain by adding topic-related word constraints to the query is that the resultant hit count is greatly diminished, canceling out any potential gain.

III.CONCLUSION

The accuracy value represents the percentage of test texts which were classified correctly by the model.

The first method: In order to verify the quality for practical use, the prototype was used for a new test set with 2,000 cases related to camera reviews. About half of the cases contained either favorable or unfavorable sentiments and the other half were neutral. System extracted sentiments for 255 cases, and 241 of them were correct in terms of the polarity of either negative or positive toward its subject within the context. Thus, without any modification of the dictionary, the prototype system achieved 94.5% (=241/255) precision with about 24% (=241/1,000) recall. [1]

The second method: A better accuracy as compared to the first method as it classifies it takes the context in consideration so The achieved efficiency is 56%.

The third method: To justify the reliability of this method 200 randomly selected candidate polar atoms in the digital camera domain. The manual classification was agreed upon in 89% of the cases and the Kappa value was 0.83, which is high enough to be considered consistent. [3]

The fourth method: In general, the addition of Osgood values does not seem to yield improvement in any of the models effectively.

In the case of the SVM with only a single Turney value, accuracy is already at 68.3% (Turney (2002) reports that simply averaging these values on the same data yields 65.8% accuracy). The Os- good values are considerably less reliable, yielding only 56.2% accuracy on their own.

IV.REFERENCES

[1]. Tetsuya Nasukawa, Jeonghee Yi: Sentiment Analysis: Capturing Favorability Using Natural Language Processing,(2003).
[2]. Namrata Godbole, Manjunath Srinivasaiah, Steven Skiena: Large-Scale Sentiment Analysis for News and Blogs.
[3]. Hiroshi Kanayama, Tetsuya Nasukawa: Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis, (2003).
[4]. Tony Mullen, Nigel Collier: Sentiment analysis using support vector machines with diverse information sources.
[5]. Thorsten Joachims. 2001. Learning to Classify Text Using Support Vector Machines. Kluwer Academic Publishers.
[6]. Chris Manning and Hinrich Schutze. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA. 1999.
[7]. C. R. Blyth. 1986. Approximate binomial confidence limits. Journal of the American Statistical Association, 81(395):843–855.
[8]. The human evaluation result for digital camera do- main (Kanayama et al., 2004).